CS 439 Final Project Report

Student Name: Abhinav Acharya

Professor: Naina Chaturvedi

Date: May 8, 2025

Project Title: Predicting Mental Health Risk Among Tech Employees Using Workplace and Lifestyle Data

# 1. Introduction

This project addresses a timely and socially significant issue: the prediction of mental health risk among tech employees. In the high-pressure, fast-paced tech industry, burnout, anxiety, and depression are increasingly common. The goal of this supervised learning project is to build a predictive model that identifies employees likely to require mental health treatment, based on factors like age, gender, employment context, benefits, and support systems.

# 2. Motivation and Background

Mental health awareness is critical in workplaces, especially in the post-pandemic era. May is also Mental Health Awareness Month, making this topic personally and academically relevant. Current research often focuses narrowly on individual predictors; this project explores a multi-factor model using behavioral, demographic, and workplace features together.

# 3. Dataset and Preprocessing

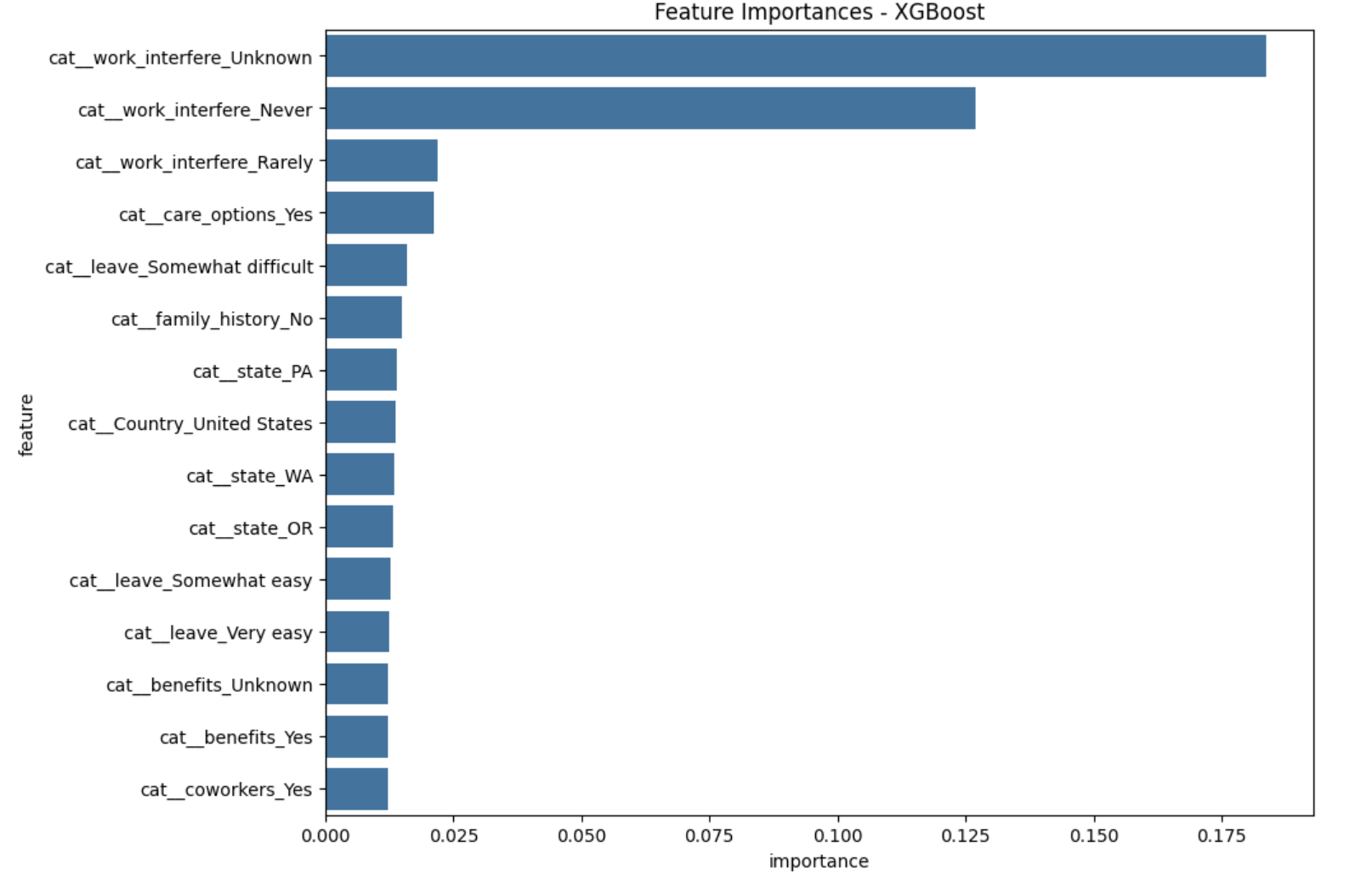
The dataset is sourced from the Open Sourcing Mental Illness (OSMI) Mental Health in Tech Survey on Kaggle, containing 1,259 responses. After data cleaning, outliers in age were removed and missing values filled. Gender values were standardized into 'Male', 'Female', and 'Other'. Columns with significant missing values or irrelevant content, like 'comments' and 'Timestamp', were dropped. Categorical variables such as 'benefits', 'leave', and 'work\_interfere' were cleaned of ambiguous responses like 'Don't know'.

After preprocessing, the final dataset had 1,251 records and 24 features. Numerical features were scaled, and categorical features were one-hot encoded, resulting in a transformed dataset with 158 total features.

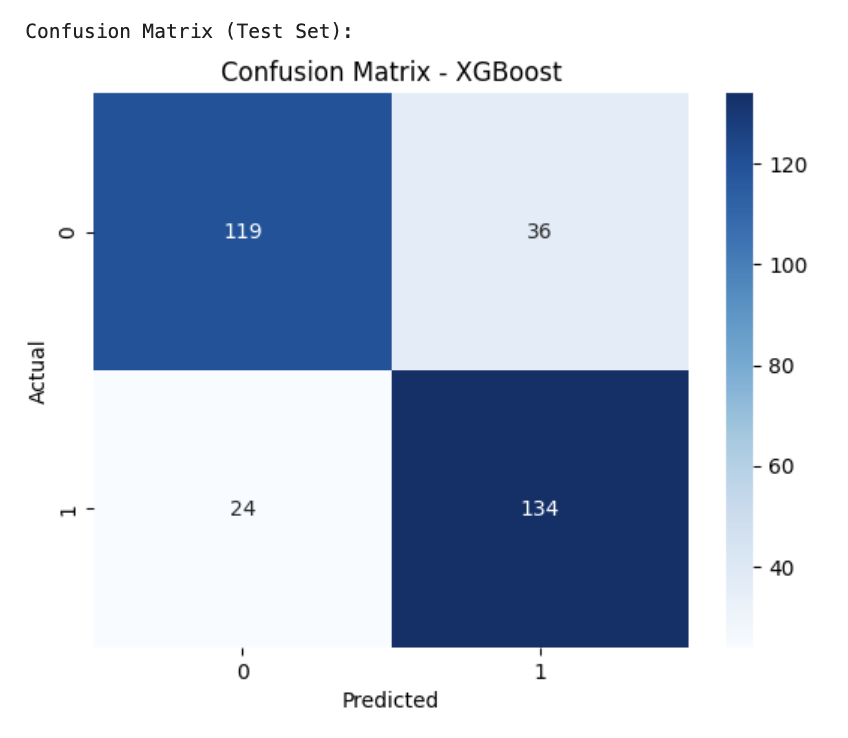
# 4. Modeling Techniques

Three classification models were trained using pipelines that included preprocessing and model fitting: Logistic Regression, Random Forest, and XGBoost. The target variable was whether the individual had sought or needed mental health treatment. To address class imbalance, all models used class weighting or appropriate scale\_pos\_weight settings.

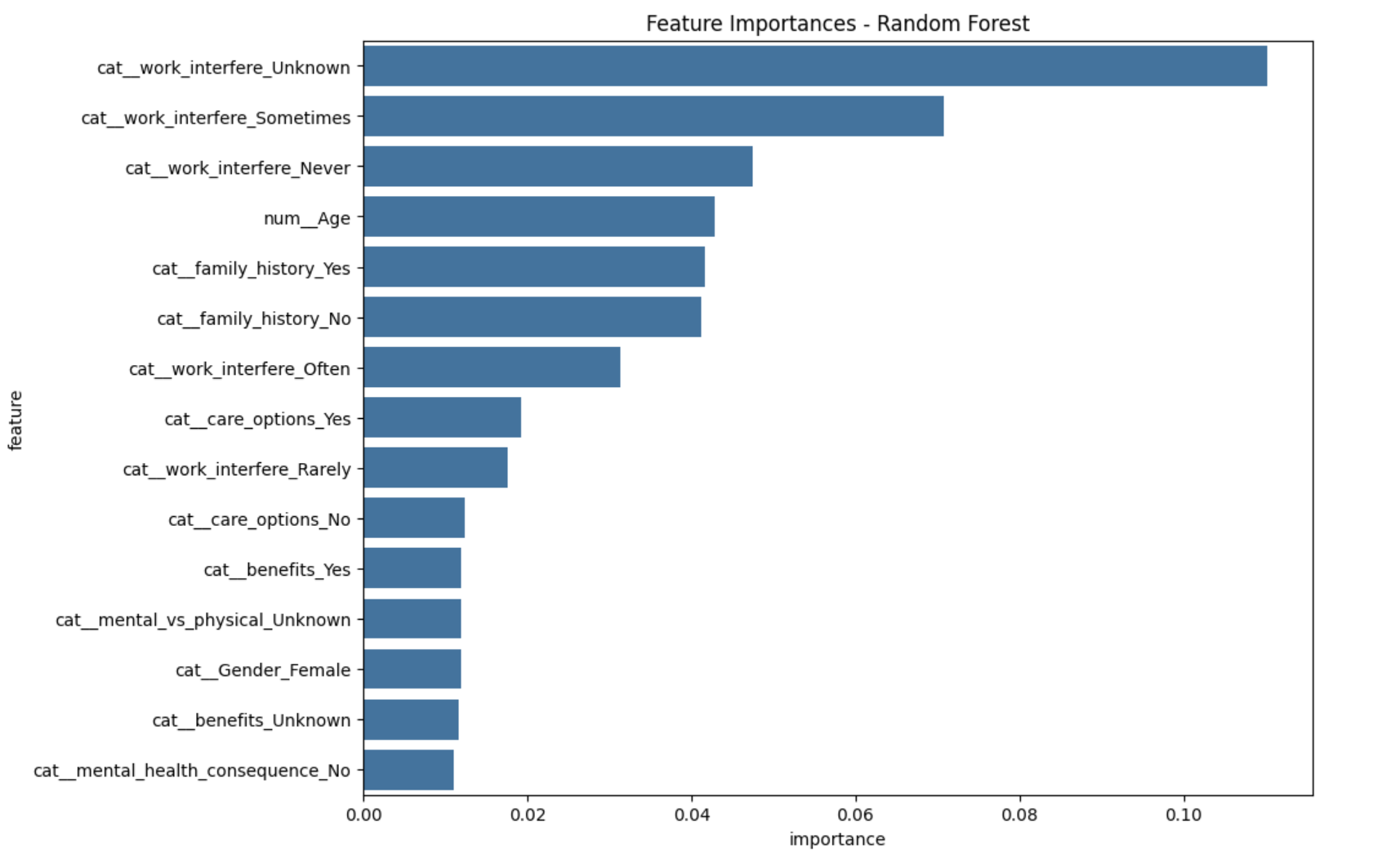
# 5. Results and Evaluation



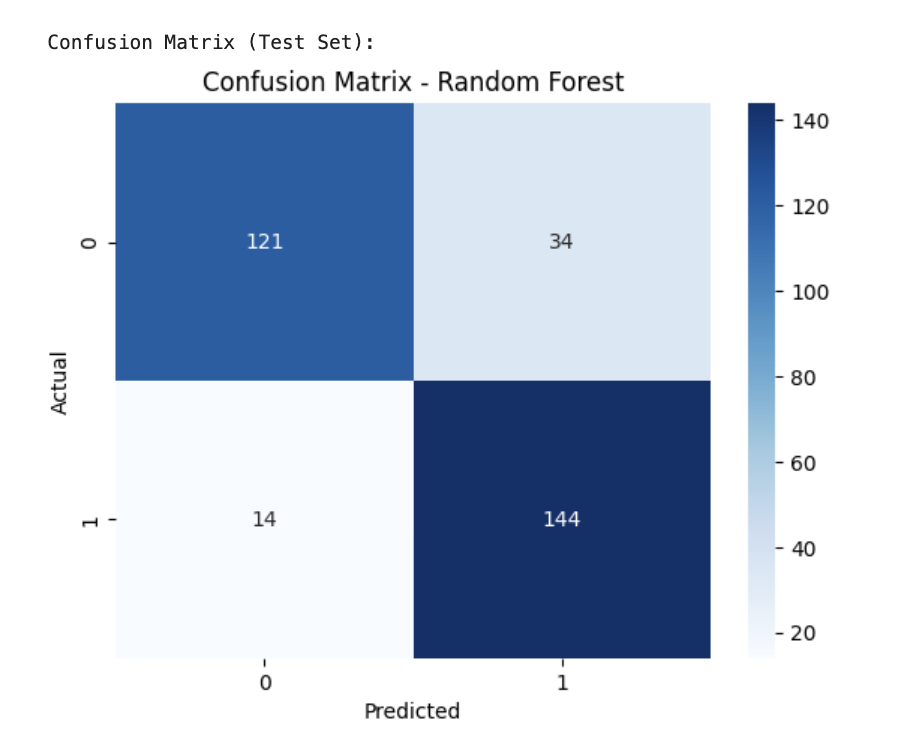
*Confusion Matrix - Logistic Regression*



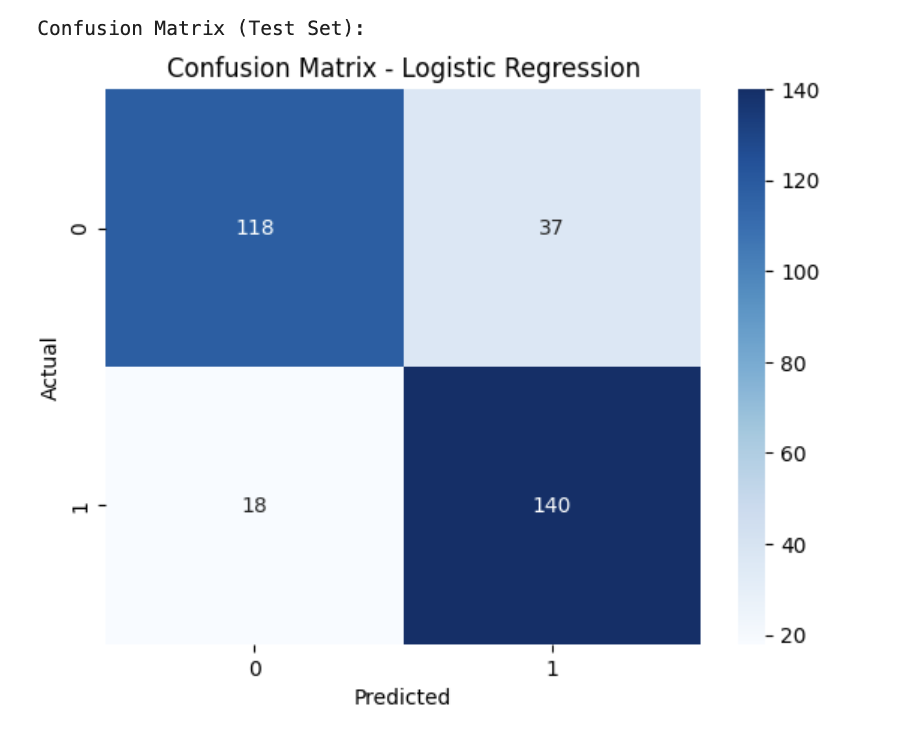
*Confusion Matrix - Random Forest*



*Confusion Matrix - XGBoost*



*Feature Importances - Random Forest*



*Feature Importances - XGBoost*

All models showed strong performance, with Random Forest achieving the highest test accuracy and recall. Recall is particularly important in this domain as it reflects the model’s ability to correctly identify at-risk individuals. Below is a summary of key metrics:

| Model | Accuracy | Precision | Recall | F1-score |  
|-------------------|-------------|-------------|---------|-------------|  
| Logistic Regression| 0.82 | 0.79 | 0.89 | 0.84 |  
| Random Forest | 0.85 | 0.81 | 0.91 | 0.86 |  
| XGBoost | 0.81 | 0.79 | 0.85 | 0.82 |

# 6. Cross-Validation and Hyperparameter Tuning

Logistic Regression underwent 5-fold cross-validation, achieving a mean recall of 0.883 and a mean F1-score of 0.837. For Random Forest, a grid search over n\_estimators and max\_depth showed the best recall with 50 trees and a depth of 10.

# 7. Discussion

Random Forest and Logistic Regression both performed well, with Random Forest slightly outperforming others in recall. The most influential predictors across models were work\_interfere, care\_options, and family\_history. These insights could guide HR policies and outreach strategies to prioritize support for employees showing risk signs.

# 8. Conclusion

This project demonstrated the feasibility of using supervised machine learning to predict mental health treatment needs among tech employees. With solid preprocessing, thoughtful model tuning, and clear visualizations, the models built here could form the basis of future research or real-world applications in employee support systems.  
  
  
GitHub Repository:  
<https://github.com/abhicommands/Final-Project-Data-Science>